ENCYCLOPEDIA OF BIOMETRICS

TITLE OF ENTRY
Speaker Databases and Evaluation

BYLINE
Alvin F. Martin
National Institute of Standards and Technology
Gaithersburg, Maryland, USA

SYNONYMS
Corpus (plural “corpora”): alternative term for database widely used to describe structured collections of speech or language data

SRE: Speaker Recognition Evaluation, frequently used to refer to the NIST SRE’s

DEFINITION
Speaker Corpora: Alternative term for speaker databases, see Speaker Databases and Evaluation

NIST SRE’s (Speaker Recognition Evaluations): Evaluations of speaker recognition systems coordinated by the National Institute of Standards and Technology (NIST) in Gaithersburg, MD, USA, 1996- 2008, see Speaker Databases and Evaluation

DET curves: Detection Error Tradeoff curves are ROC (Receiver Operating Characteristic) type curves showing the range of operating points of systems performing detection tasks as a threshold is varied to alter the miss and false alarm rates and plotted using a normal deviate scale for each axis. DET curves have the property that if the underlying score distributions for the two types of trials are normal, the curve becomes a straight line. They have been widely used to present the performance characteristics of speaker recognition systems. See [1] and Speaker Databases and Evaluation

[1] Martin, A. F., et al., “The DET Curve in Assessment of Detection Task Performance”, *Proc. Eurospeech ‘97,* Rhodes, Greece, September 1997, Vol. 4, pp. 1899-1903

MAIN BODY TEXT
Introduction

Expanding interest in the use of biometrics for security purposes has brought increasing attention to the use of speech as a biometric. Speech fits naturally into the list of likely biometric modalities. It is an activity engaged in by essentially everyone, and is one of the primary means by which people identify those they know.

But speaker recognition has not heretofore been seen as among the most useful biometrics for general security applications. There has been much more development effort on the use of face,

fingerprint, and iris. Recognition of speakers by voice has been seen as more of a niche application, largely because of the special difficulties associated with the collection of quality speech input, and perhaps because of a particular advantage it may offer.

In this introduction we briefly discuss some key issues related to speaker recognition as a biometric. In the following section we discuss some of main databases that have been used for speaker recognition research and evaluation. In the final section we discuss the leading technology evaluations of speaker recognition that have been conducted and are ongoing.

Speaker recognition may be divided into speaker identification (many to one decision) and speaker verification or speaker detection (one to one decision). Due perhaps because to performance limitations and the strategic decision in the NIST evaluations to make speaker detection the core task, the research community has in recent years come to focus on the latter. This better represents the areas of current practical applications and, of course, ultimately superior performance for the latter would make possible the former.

Defining a "standard" test for speaker matching is not simple. Numerous environmental factors affect the quality of any voice signal collected, and these may, depending on the collection configuration and circumstances, be very difficult to control. There are many choices of protocol to be made, involving in particular the type of speech and specific words, as well as the amount of speech, to be collected. These issues are very much application dependent, and operational consensus is very hard to achieve.

Best performance in voice recognition is achieved when a consistent wideband high quality audio channel is available for all speech input. But the needed quiet room environment can be expensive and often impractical to set up, and may be rather demanding on the user in terms of speaking into a close talking microphone. Meanwhile, competing biometrics may more easily provide similar capability.

The particular advantage offered by voice as a biometric is that it is transmissible over telephone channels. Telephone handsets, landline or cellular, are ubiquitous in modern society. The variability of telephone handsets and telephone channels makes the recognition task far more difficult and degrades the quality of performance. Nevertheless this has been the area of greatest application interest, and thus of greatest interest for evaluation.

One key distinction among speaker recognition applications is the type of involvement of the speaker in the process. The speaker may or may not be aware of the recognition process, and if aware, may or may not seek to cooperate with it.

Applications involving access, whether to a physical location or to information, are likely to involve cooperative and motivated users. The system can then prompt the speaker to say a specific phrase, or even a previously agreed upon passphrase (perhaps an account number), allowing the recognition to be text-dependent and even combined with a pin number for greater effective performance. Commercial applications often rely on the use of short phrases spoken by cooperative users, with the system's knowledge of what is to be said (text-dependence) helping to aid performance despite the limited amount of speech involved and the difficulties posed by variable telephone channel conditions.

Forensic applications, on the other hand, will involve either an unaware or uncooperative user, and other applications will involve listening in on unaware speakers. Here text-dependent recognition is not an option. A characteristic of this type of application, however, is that it may be possible to collect rather long durations of speech from the speakers, whereas a cooperative scenario requires that valid speakers be able to enroll and obtain access after brief speaking intervals. This can allow systems to learn more about a target speaker's speaking style and idiosyncrasies. The frequency of occurrence of particular words and phrases in someone's natural (determined with the aid of automatic speech recognition technology for word transcription) may powerfully aid recognition performance.

Databases

The era of standard corpora (or databases) for speech processing applications began in the mid-1980's as modest priced computers became capable of performing the necessary signal processing and the costs of storage media fell significantly. The Speech Group at NIST (National Institute of Standards and Technology) played an early role in making corpora of interest available at reasonable cost in CD-ROM format. Since its founding in 1992, the Linguistic Data Consortium (LDC) at the University of Pennsylvania has been the primary repository of speech corpora in the United States. (ELRA, the European Language Resources Association, plays a similar role in Europe.) The corpora described here are available through the LDC and are described in its online catalog (www.ldc.upenn.edu/catalog).

There are particular properties needed of corpora to support speaker recognition research. A substantial number of different speakers must be included, and most particularly, there needs to be a number of different recorded sessions of each speaker. Applications require speakers to enroll in the system at one time and to be successfully detectable at a later time. Multiple recording sessions, particularly when recorded over time varying telephone channels, are essential to represent this. Moreover, telephone handsets vary, so it is desirable, for most real-world applications, to have different sessions using different handsets. It has been seen that recognition performance over the telephone is considerably better if speakers can use the same handset in both training (enrollment) and test. This is particularly so if impostor speakers use different handsets from speakers of interest, as is inherently the case in most collection protocols. Otherwise, systems may be doing channel recognition rather than speaker recognition. Thus a corpus such as Macrophone (the U.S. contribution to the international Polyphone corpus) , collected to support multiple types of speech research and containing telephone speech of a variety of types from a large number of speakers, has been of limited usefulness for speaker recognition because of having only a single session for each speaker.

One early corpus widely used for speaker research was TIMIT, produced from a joint effort by Texas Instruments (TI) and the Massachusetts Institute of Technology (MIT), along with SRI International, with sponsorship by DARPA (Defense Advanced Research Projects Agency). TIMIT is a corpus of read speech, containing 10 phonetically diverse sentences spoken by each of 630 speakers chosen to represent 8 major dialect regions of the United States. Its basic implementation consists of high quality microphone speech, but versions of the data sent through a lower quality microphone channel or different types of telephone channels were also produced.

TIMIT was collected for multiple types of speech processing, but was very popular in speaker identification/recognition research through much of the 1990's, partly because few alternatives were widely available and partly because its limited vocabulary and high recording quality supported the attainment of impressive text-dependent performance results. It was a source of some frustration to leading researchers at speaker recognition workshops held in the 1990's that paper after paper discussed systems performance on TIMIT, rather than on any "real" data.

An early corpus collected specifically for speaker recognition was the KING Corpus. It involved 51 male speakers from whom 10 sessions of about 30 seconds each were collected. The speech was collected simultaneously over a wideband channel and a narrowband telephone channel. There were 25 speakers whose speech was collected in New Jersey, and 26 whose speech was collected in San Diego. For the San Diego speakers, researchers attempting to do speaker detection noted that there was a "great divide" between the first five and second five of the ten sessions involving narrowband speech due to replacement of a circuit component during the collection. The spectral slope characteristics turned out to be very different on the two sides of this divide, although the collectors never noticed it. Much effort was devoted to understanding and coping with this phenomenon, and this led to greater awareness of the effects of channel

characteristics for speaker recognition using telephone speech, and considerable later research effort to compensate for such channel differences.

A third early corpus for text-dependent recognition of high quality speech was known as the YOHO Corpus. It was collected (like KING) by ITT under a US government contract in 1989. There were 138 speakers who each had 4 enrollment and 10 verification sessions. Each session involved speaking "combination locks" each consisting of three two digit numbers. There were 24 spoken phrases in enrollment sessions, and 4 in verification sessions. This was clearly intended for access applications involving cooperative speakers.

It does not appear that these early corpora were used in multi-site evaluation, but were used extensively in evaluating individual site research projects. As will be noted, it has been difficult to find sufficient interest and agreement on protocols for text-dependent evaluation in the speaker arena. Table 1 summarizes these early corpora.

The modern era in the collection of corpora for speaker recognition perhaps began with the collection of the Switchboard Corpus for DARPA by TI in the early 1990's. This collection of about 2400 two-sided telephone conversations from over 500 participating speakers was originally intended for multiple purposes, including word spotting, topic spotting, and speaker spotting in the terminology used at the time. An automatic system was created which allowed registered participants to call in at specified times to a "robot operator" which attempted to contact other registered participants and initiate a two-way conversation on one of about 55 pre-specified topics the participants had indicated would be acceptable. Thus the conversants generally engaged in an at least somewhat serious discussion for five minutes or more with someone whom they did not know. A speaker's topic and conversational partner were in general never repeated in different conversations. A subset of the participating speakers were encouraged to make a sizable (double-digit) number of different conversations and to use multiple telephone handsets in them.

Switchboard-1 (so denoted when similar corpora followed) was used in a couple of limited U.S. government sponsored evaluation of speaker spotting (and a similar evaluation of topic spotting) in the early 90's, but it proved to be a popular corpus for further study and research. Somewhat surprisingly it was used in subsequent years for general evaluation of automatic speech (word) recognition, as the focus of such evaluation shifted to natural unconstrained conversational speech. In 1996 it provided the data for the first of the series of NIST Speaker Recognition Evaluations (SRE's) discussed below. A subset of 40 of the most prolific corpus speakers was used as the target speaker set in this evaluation.

The success of Switchboard-1, particularly for speaker recognition, led to the collection of the multi-part Switchboard-2 and Switchboard Cellular Corpora. Each involved hundreds of speakers taking part in a number of different conversations using multiple telephone handsets. This was important as the early NIST evaluations established that telephone handset variation between training and test very much affected system performance, and the desire was to truly recognize speakers and not merely handsets. The Switchboard-2 Corpora each concentrated largely on speakers from a specific area of the United States, relying mainly on college or early post-college age students. Switchboard Cellular was collected in light of the increasing use of cellular telephone handsets in the United States.

The Switchboard Corpora supplied the bulk of the evaluation data used for the annual NIST evaluations from 1996 to 2003. Table 2 summarizes these corpora.

Around 2003 the LDC moved to a somewhat different collection model from that used in the Switchboard Corpora. The "Fisher" platform was similar to that used for Switchboard, but it could also initiate a search for paired conversants without one party initiating matters with a call into the system. It was to prove useful in new corpus collections for general speech recognition and for language recognition, but was also applied to speaker recognition collection. For this

purpose the multi-part Mixer Corpus has been collected. It was used in the 2004, 2005, and 2006 SRE's, and will be used in the 2008 SRE.

The Mixer collections have expanded the types of speaker data collected in two major ways. The first is the inclusion of conversations in multiple languages. LDC recruited a sizable number of bilingual speakers (with English as one language) and utilized the collection protocol to pair up speakers of a non-English language, who received a bonus for talking in their other language. It became feasible, for example, to have certain specified days devoted to the collection of calls in specified languages. This supported investigation of the effect of language, and of language change between training and test, in speaker recognition performance.

Second, the Mixer corpora have included some conversations in which participants were recorded simultaneously over the telephone and over eight or more different microphones. These included a range of close talking, near-field, and far-field microphones to support comparison of performance over different microphone types, and the examination of cross channel condition differences between training and test. This was accomplished by having select groups of participants come to a special room at two collection sites where all of the microphones could be carefully placed while they used provided cell phones to call the automatic system and be paired with participants in the usual way.

The Mixer 5 Corpus collected in 2007 contains a further variation on this theme. Its 300 speakers each participated in a series of six structured "interviews" of about a half hour each, occurring over at least three different days. The bulk of each interview involves conversational speech, but with an interviewer who is present in the room and provides appropriate prompts. The subject's speech is recorded over a dozen or so carefully placed microphones, but not over a telephone line. Over the course of the six sessions the subject gets to know the interviewer, and this changes the nature of the spoken dialog. Each interview also contains a brief period of standard repeating questions, and periods of different types of read speech. Each participant also makes two simulated phone calls where side tones are used to encourage each rather high or rather low vocal effort. Each interview subject is also paired in the usual way in about ten regular phone conversations with unknown interlocutors outside of the interviews. This data will be used in upcoming NIST SRE's and may offer some interesting contrasts with previous results.

The Mixer Corpora are discussed further in [1, 2, 3]. Table 3 summarizes the Mixer Corpora.

Evaluations

Evaluations of speaker recognition require a sponsor or sponsors and participants. Sponsors must be willing to commit the necessary resources to support an evaluation infrastructure. Most important, they must support the collection of speech databases appropriate to the tasks or applications of interest to them, and thus suitable for the particular evaluation.

Participants must be willing to take part in evaluation, to discuss the systems they develop, and to have their performance results presented to the evaluation community. They must be ready to do this not knowing in advance whether their evaluation performance will compare favorably or unfavorably with that of the other participants.

The most notable series of evaluations of recent years have been those coordinated by the National Institute of Standards and Technology (NIST), an agency of the U.S. Department of Commerce, in Gaithersburg, Maryland, USA. The NIST evaluations have received sponsorship support and guidance from interested U.S. government agencies involved in defense, intelligence and law enforcement.

There were a couple of preliminary evaluations held in 1992 and 1995, each utilizing a limited number of target speakers from the Switchboard-1 Corpus. They did not involve the scoring

metric of the later evaluations, described below, and looked at the range of operating points (receiver operating characteristic curves) of each target speaker separately rather than combining them based on a required calibration threshold into a single curve as will be described below. The 1995 evaluation was the first to analyze and note the effect on performance of having a speaker's training and test segments come from the same or different telephone numbers, and thus same or different telephone handsets. These evaluations each had only about a half dozen participants, mainly from the United States.

The NIST evaluations assumed basically their present form in 1996, and were conducted annually from 1996 to 2006, with the next one set to occur in 2008. These have all included as the core task text-independent speaker detection in the context of conversational telephone speech. The 1996 evaluation selected 40 of the more prolific Switchboard-1 speakers as target talkers, and used other corpus speakers for non-target trials. The subsequent evaluations have all utilized hundreds of speakers from the LDC corpora involved (Switchboard through 2003, Mixer subsequently), and have followed the practice of allowing the target speakers to also serve as impostor speakers for non-target trials. The evaluation plan documents and other information related to these evaluations may be found at http://www.nist.gov/speech/tests/sre/index.html.

Participation in the NIST speaker recognition evaluations has grown steadily and become worldwide in scope. The number of participating sites has grown to reach approximately 35 in 2006.The numbers of participants noticeably increased in 2002 and subsequent years, perhaps because of growing interest in biometric technologies after the events of 2001.

Of the growing number of participants in recent years, only about half a dozen have been sites in the United States, with a majority in Europe, and an increasing number from the Far East. The greatest numbers of participants have been from the U.S., France, and China. Other participants have been from Canada, various European countries, Singapore, Australia, Israel, and South Africa.

Most of the sites participating in the NIST evaluations have been from academic institutions. Some government funded research institutions or companies involved in government research have also participated. Not frequently represented, however, have been smaller commercially oriented companies. This may be due in part to the text-independent and research, rather than application, oriented type of evaluation being conducted, but also bespeaks a reluctance to participate in evaluations where competitors may show superior performance results.

Evaluation requires a performance measure. For detection tasks there are inherently two types of error. There are trials where the target is present (target trials) but a "false" decision is made by a system. Such errors are misses. And there are trials where the target is not present (non-target or impostor trials) but a "true" decision is made. These are referred to as false alarms. Thus we may speak of a miss rate for target trials and a false alarm rate for non-target trials.

The NIST evaluations have used a linear combination of these two error rates as its primary evaluation metric. A decision cost function (DCF) is defined as

$$\begin{aligned} DCF &= C_{Miss} \times P_{Miss|Target} \times P_{Target} \\ &+ C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1-P_{Target}) \end{aligned}$$

where C_{Miss} represents the cost of a miss, C_{FA} the cost of a false alarm, and P_{Target} the prior probability of a target trial. These are three somewhat arbitrary, and certainly application dependent, parameters. The NIST evaluations have used parameter values

C_{Miss}	$C_{FalseAlarm}$	P_{Target}
10	1	0.01

These have been viewed as reasonable parameters for applications involving an unaware user, where most speech segments examined are likely to be of someone other than the target of interest, but where detecting instances of the target have considerable value. Note that P_{Target} need not represent the actual target richness of the evaluation trials, but may be chosen based on possible applications of interest. The NIST evaluations have generally had an approximately ten to one ratio of non-target to target trials, to minimize the variance of the metric in light of the parameter values chosen.

A detection task inherently involves two types of error, and a system may be expected to be able to tune its performance to vary the relative frequency of the two error types. In the NIST evaluations systems have been required for each trial to produce not only a decision, but also a score, where higher scores indicate greater likelihood that the correct decision is "true". A decision threshold may then be varied based on this score to show different possible operating points or tradeoffs between the two types of error. Note that the evaluations have required that this threshold be the same for all target speakers.

The most informative way of presenting system performance in the NIST SRE's has been to draw a curve showing the operating points and the tradeoff in the error rates. This is easily done by varying the decision threshold based on the scores provided. A simple linear plot is know as an ROC (Receiver Operator Characteristic) curve, but a clearer presentation is obtained by putting both error rates on a normal deviate scale to produce what NIST has denoted a DET (Detection Error Tradeoff) curve [4]. This has the nice property that if the underlying error distributions for the miss and false alarm rates are normal, the resulting curve is linear.

Figure 1 shows DET curves for the systems in the core test done in the 2006 NIST SRE. These are curves representing the performance of the primary systems submitted by over 30 sites participating in the evaluation. Better systems have performance curves closer to the lower left corner of the plot. The actual decision point of each performance curve is denoted with a triangle, and a 95% confidence box is drawn around these, while circles are used to denote the points corresponding to the minimum DCF operating points. The closer these two specially denoted points on each curve, the better the system did at calibrating its decision threshold for hard decisions. For example, for the best performing system shown, the actual decision point has a false alarm rate of about 2% and a miss rate of about 7%, while the minimum DCF point has a false alarm rate of about 1% and a miss rate of about 11%. This gives a sense of the level of current state-of-the-art performance for speaker detection on this type of telephone data.

A possible alternative non-parametric information theoretic type of metric has been proposed to be applicable to a range of applications, and has been included as an alternative measure in the most recent NIST evaluations, provided the system specifies that its likelihood scores may be viewed as log likelihood ratios. This metric is discussed in [5].

While the basic detection task has remained fixed, there have been multiple test conditions in most of the evaluations, and these conditions have varied over the years. In particular there has been variation in the durations of the training and test segments. While the earlier evaluations focused on landline phones and the varying types of telephone handsets (carbon-button vs. electrets microphone), in the new millennium there is was greater focus on the effect of cellular transmission and newer types of handsets as these became common in the U.S. Certain additional data sources, such as a small FBI forensic database and a Castilian Spanish corpus known as AHUMADA (neither one easily available) were used in one or two evaluations.

The earlier evaluations used fixed durations of speech, as determined by an automatic speech detector. Later evaluations allowed more variation in duration within each test condition.

Starting in 2001 there was greater interest in longer durations for training and test. This was largely as a result of some research suggesting that with effective word recognition, higher level lexical information about a speaker could be effectively combined with more traditional lower level acoustic information [6]. As a result of the apparent success of such an approach in the 2001evaluation, a major summer research program was carried out at Johns Hopkins University in the summer of 2002 (see http://www.clsp.jhu.edu/ws2002/groups/supersid/). Since then, "extended" training conditions, where the training consists of multiple (often eight) conversation sides have been a major part of the evaluations. The earlier NIST evaluations are described further in [7, 8, 9].

The introduction of Mixer data in 2004 inaugurated a new era in the NIST evaluations. The inclusion of calls in multiple languages and cross language trials introduced a new wrinkle that affected overall performance. The latest evaluations have also introduced test conditions involving multiple microphones and cross channel trials, that will be a major focus in 2008 and beyond. The recent SRE's are discussed in [10, 11, 12].

Have the evaluations shown progress in performance capabilities over the years? They have, but changes in the test conditions from year to year and in the types of data used have complicated performance comparisons. Figure 2 from [13] attempts to sort these matters out, and summarizes the DCF scores of the best evaluation systems across ranges of years involving more or less consistent test conditions.

The NIST SRE's have been the most notable evaluations in speaker recognition in recent years. They have concentrated on a basic speaker detection task not tied to any specific current commercial application. This has made it attractive to a large range of research sites around the world to participate in these evaluations.

One other notable evaluation in the field was conducted by TNO in the Netherlands in 2003. It featured a protocol very similar to that of the NIST evaluations, but utilized actual forensic data provided by the Dutch police. Its very interesting results are discussed in [14], but the data used was only provided to the evaluation participants for a limited time and purpose and is not otherwise available.

There have been other efforts to encourage evaluation.. Research in speaker recognition technology has been advanced by the series of Odyssey workshops. These were held in Martigny, Switzerland in 1996, Avignon, France in 1998, Crete, Greece in 2001 (where the name "Odyssey" was adopted), Toledo, Spain in 2004, San Juan, Puerto Rico in 2006, and Stellenbosch, South Africa in 2008. For the 2001 workshop an evaluation track was included. This included both a text-independent track based on the preceding NIST evaluation, and a text-dependent track. Participation, particularly in the text-dependent track, was very limited, perhaps demonstrating the difficulty of persuading companies or organizations to participate in this inherently application-specific and more immediately commercially oriented field.

The European Union has sponsored a multi-year program to develop biometric technologies denoted BioSecure (http://www.biosecure.info/), with speaker as one of the included technologies. Evaluation is intended to be part of this program, in particular including evaluation of the fusion of multiple biometrics. As of 2007, however, speaker recognition evaluation appears not to have begun.

The NIST evaluations will resume in 2008, and may be held in alternate years in the future. They will feature an increased emphasis on cross channel recognition. Whereas in 2005 and 2006 the core test involved only telephone speech, with cross channel (train on telephone, test on microphone) an optional additional test, the core test condition is expected to require processing of a mix of training or test segments including both telephone and microphone speech, with some of the trials including different channels in training and test. This will utilize at least both Mixer 3 and Mixer 5 type data. Evaluation performance, however, will be

subsequently analyzed to distinguish performance on telephone, microphone, and cross-channel trials. A number of different microphone types from the Mixer 5 data will be included.

REFERENCES

[1] Christopher Cieri, Joseph P. Campbell, Hirotaka Nakasone, David Miller, Kevin Walker, *The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data,* LREC 2004: Fourth International Conference on Language Resources and Evaluation, Lisbon

[2] Christopher Cieri, Walt Andrews, Joseph P. Campbell, George Doddington, Jack Godfrey, Shudong Huang, Mark Liberman, Alvin Martin, Hirotaka Nakasone, Mark Przybocki, Kevin Walker, *The Mixer and Transcript Reading Corpora: Resources for Multilingual, Crosschannel Speaker Recognition Research,* LREC 2006: Fifth International Conference on Language Resources and Evaluation

[3] Christopher Cieri, Linda Corson, David Graff, Kevin Walker, *Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora***,** Interspeech 2007, Antwerp, August 2007

[4] Martin, A. F., et al., "The DET Curve in Assessment of Detection Task Performance", *Proc. Eurospeech '97,* Rhodes, Greece, September 1997, Vol. 4, pp. 1899-1903

[5] Brummer, N., and du Preez, J, "Application-independent evaluation of speaker detection" in *Computer Speech & Language*, volume 20, issues 2-3, April-July 2006, pages 230-275

[6] Doddington, G., "Speaker Recognition Based on Idiolectal Differences Between Speakers", *Proc. Eurospeech '01,* Aalborg, Denmark, September 2001, Vol. 4, pp. 2521-2524

[7] Martin, A. F. and Przybocki, M. A.,, "The NIST Speaker Recognition Evaluations: 1996-2001", *Proc 2001: A Speaker Odyssey*, Chainia, Crete, Greece, June 2001, pp. 39-43

[8] Martin, A. F., Przybocki, M. A., and Campbell, J. P., "The NIST speaker recognition evaluation program", in Wayman, J. et al., editors, *Biometric Systems: Technology, Design and Performance Evaluation*, ch. 8, pp. 241-262, Springer, 2005

[9] Przybocki, M. A. and Martin, A. F., "NIST Speaker Recognition Evaluation Chronicles", *Proc. Odyssey 2004: The Speaker and Language Recognition Workshop,* Toledo, Spain, June 2004

[10] Przybocki, , M. A., Martin, A. F., and Le, A. N., "NIST Speaker Recognition Evaluation Chronicles – Part 2", *Proc. Odyssey 2006: The Speaker and Language Recognition Workshop*, San Juan, PR, June 2006

[11] Przybocki, M. A., Martin, A. F., and Le, A. N., "NIST Speaker Recognition Evaluations Utililizing the Mixer Corpora – 2004, 2005, 2006", *IEEE Transactions on Audio, Speech, and Language Processing*, V. 15, N. 7, September 2007

[12] Martin, A. F., "Evaluations of Automatic Speaker Classification Systems" in, Muller, C (ed.) *Speaker Classification I*, Springer, 2007, pp. 313-329

[13] Reynolds, D. A., Keynote talk "Speaker and Language Recognition: A Guided Safari", *Proc. Odyssey 2008: The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, January 2008

[14] van Leeuwen , D. A. et al., "NIST and NFI-TNO evaluations of automatic speaker recognition", *Computer Speech & Language*, Vol. 20, Issues 2-3 , April-July 2006, pp. 128-158

Figure 1: DET (Detection Error Tradeoff) Curves for the primary systems of participating sites on the core test of the 2006 NIST SRE.

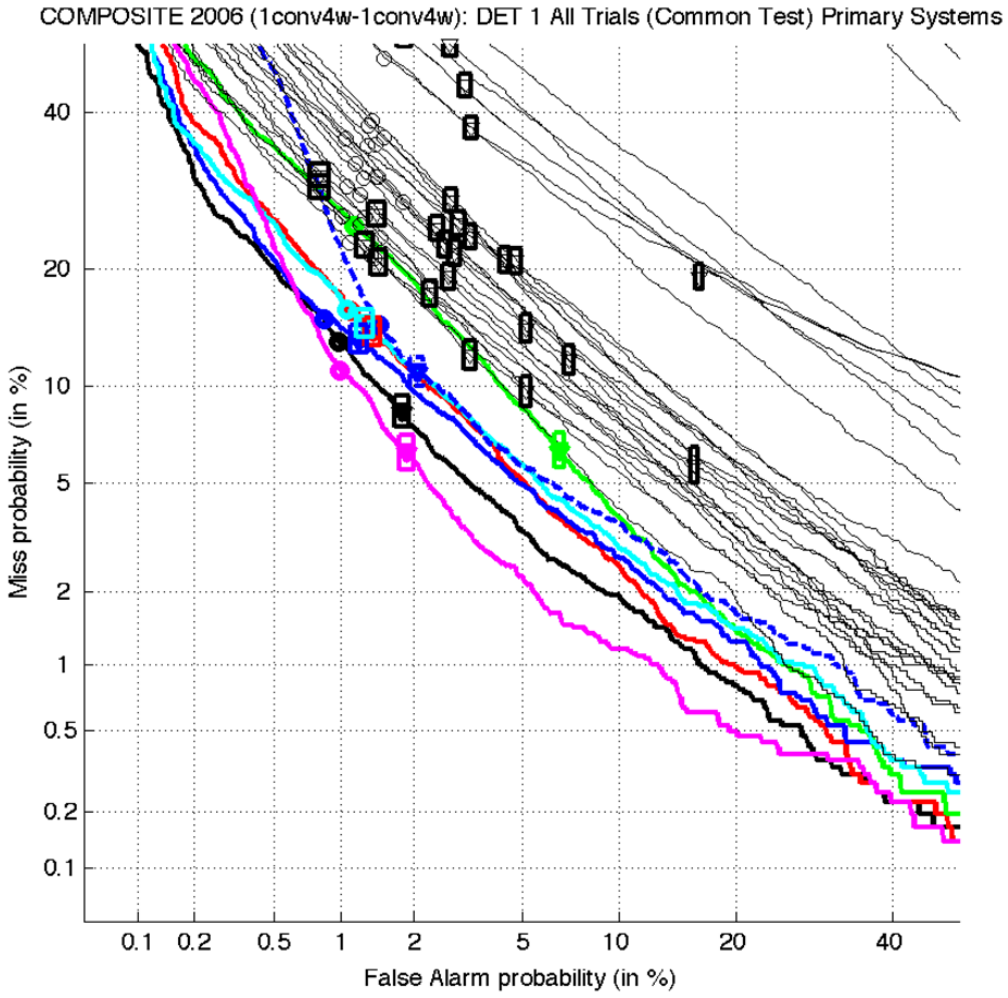

Figure 2: DCF (Decision Cost Function) values for the best (lowest DCF) systems on different roughly comparable evaluation conditions over multiple years during the course of the NIST SRE's from 1996 to 2006.

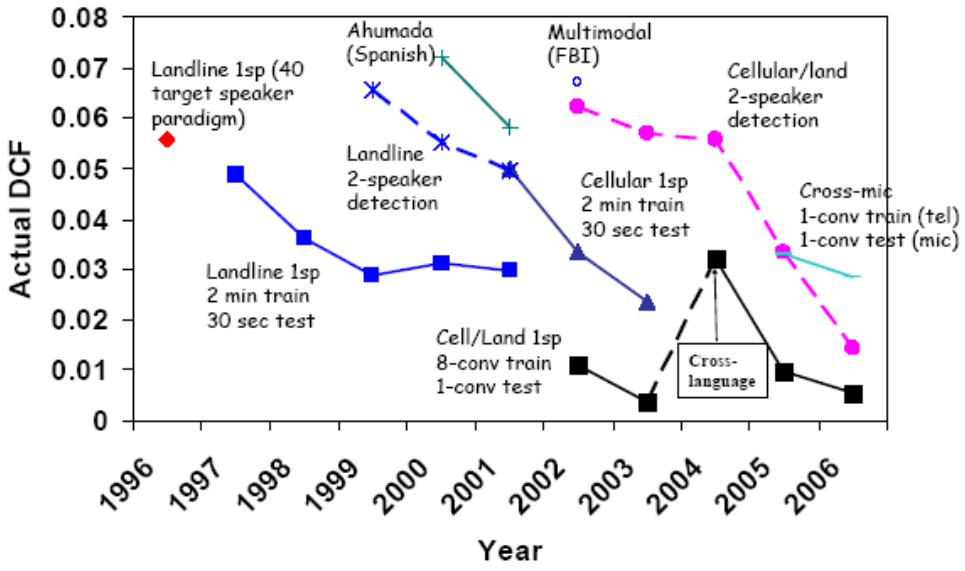

Table 1: Some early corpora used for speaker recognition

Year	Corpus	Size	Types of Speech
Early 1980's	TIMIT	630 speakers of eight major US English dialects, 10 sentences each; alternative versions run original wideband data through other specified channels	Read speech of phonetically rich sentences
1987	KING	51 male speakers (25 New Jersey, 26 San Diego), 10 sessions each recorded on both a wide-band and a narrow-band channel	Sessions contain 30 seconds of speech on an assigned topic
1989	YOHO	138 speakers with 4 enrollment sessions (24 phrases) and 10 test sessions (4 phrases)	"Combination lock" phrases

Table 2: The Switchboard Corpora; collection years are approximate

Year	Corpus	Size	Types of Speech
1990/ 1991	SWBD I	543 speakers, 2400 two-sided conversations	USA conversational telephone speech on assigned topics
1996	SWBD II phase 1	657 speakers, 3638 conversations	Primarily US Mid-Atlantic, conversational telephone
1997	SWBD II phase 2	679 speakers, 4472 conversations	Primarily US Mid-West, conversational telephone
1997/ 1998	SWBD II phase 3	640 speakers, 2728 conversations	Primarily US South, conversational telephone
1999/ 2000	SWBD cellular p1	254 speakers, 1309 conversations	Primarily cellular GSM, USA conversational
2000	SWBD cellular p2	419 speakers, 2020 conversations	Cellular, largely CDMA, USA conversational

Table 3: The Mixer Corpora; collection years are approximate

Year	Corpus	Size	Types of Speech
2003	MIXER p1 and p2	600 speakers with 10 or more calls 200 with 4 cross-channel calls	Conversational, some calls in 4 non-English languages
2005	MIXER p3	1867 speakers with 15 or more calls	Conversational, includes calls in 19 languages
2007	MIXER p4	200 speakers making 10 calls including 4 cross-channel	Conversational, primarily English
2007	MIXER p5	300 speakers doing 6 interviews and generally 10 phone calls	Conversational in interview setting, some read speech